Methods for Characterizing and Comparing Populations of Shock Wave Curves

Curtis B. Storlie, Michael L. Fugate, David M. Higdon, Aparna V. Huzurbazar
Elizabeth G. Francois, Douglas C. McHugh

Los Alamos National Laboratory

Date: June 3, 2012

Abstract

At Los Alamos National Laboratory, engineers conduct experiments to evaluate how well detonators and high explosives work. The experimental unit, often called an "onionskin", is a hemisphere consisting of a detonator and a booster pellet surrounded by high explosive material. When the detonator explodes a streak camera mounted above the pole of the hemisphere records when the shock wave arrives at the surface. The output from the camera is a two-dimensional image that is transformed into curve that shows the arrival time as a function of polar angle. The statistical challenge is to characterize the population of arrival time curves, and to compare the baseline population of onionskins to a new population. The engineering goal is to manufacture a new population of onionskins that generate arrival time curves with the same shape as the baseline. We present two statistical approaches which test for differences in mean curves, and provide simultaneous confidence bands for the difference: (i) a B-Spline basis approach, (ii) a Bayesian hierarchical Gaussian process approach. In problems that involve complex modeling with modest sample sizes, it is important to apply multiple approaches with complementary strengths such as these to determine whether all approaches provide similar results. The performances of the two approaches are compared on several simulations that were constructed to mimic the actual onionskin analysis. Finally an analysis of historical onionskin data is presented.

Keywords: Nonparametric Regression, B-Splines, Gaussian Process, Hierarchical Modeling, Functional Data Analysis, Onionskin.

Running title: Onionskin Analysis

1 Introduction

At Los Alamos National Laboratory, engineers conduct experiments to evaluate how well detonators and high explosives work. The experimental unit, often called an "onionskin" (Dallman 1988, Hogan, Adams & Alrick, K., et. al. 1999, Hooks, Morris, Hill & Francois 2008)

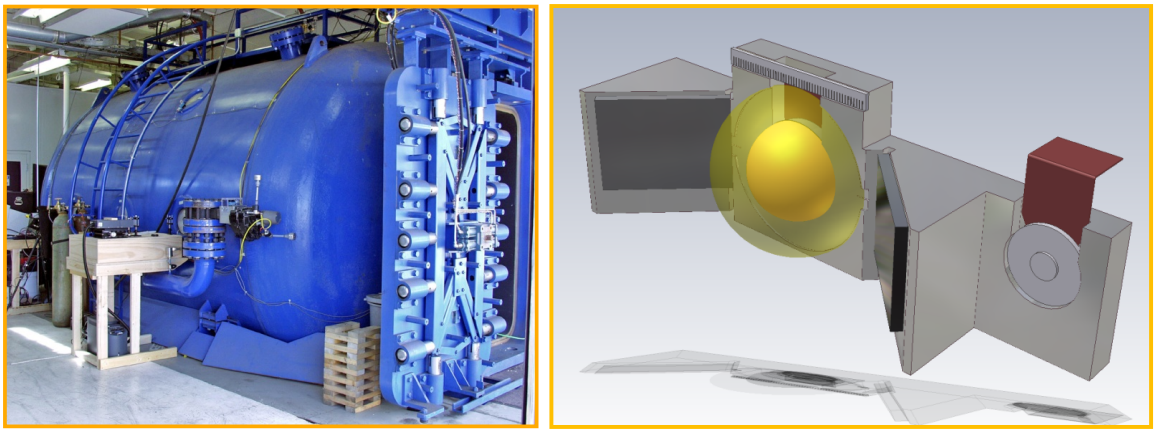
or "snowball" (Lundberg 1996), is a hemisphere consisting of a detonator and a booster pellet surrounded by high explosive material with the center of the detonator and pellet located directly under the pole at the equator of the hemisphere. When the detonator explodes a streak camera mounted above the pole records when the shock wave arrives at the surface. The output from the camera is a two-dimensional image that is transformed into a discretized curve that shows the arrival time as a function of polar angle; an arrival time curves is often referred to as onionskin curve.

Figure 1 shows a drawing of an onionskin mounted in a test device. Above the device is a streak camera image recorded from an actual experiment. Also shown are nine onionskin curves from historical tests performed at Los Alamos that, for purposes of this paper, are considered the baseline data. Each discretized curve gives the arrival time in microseconds (μs) of the shock wave on the surface of the onionskin as a function of polar angle. The arrival times for different curves are observed at different angles. For reference, the sample size and range of observed angles for each curve in Figure 1 is given in Table 1. Although observed angles can range between $\pm 90°$, engineers are only interested in the shape of the curves between $\pm 80°$. Consequently, the analysis below excludes data outside this range.

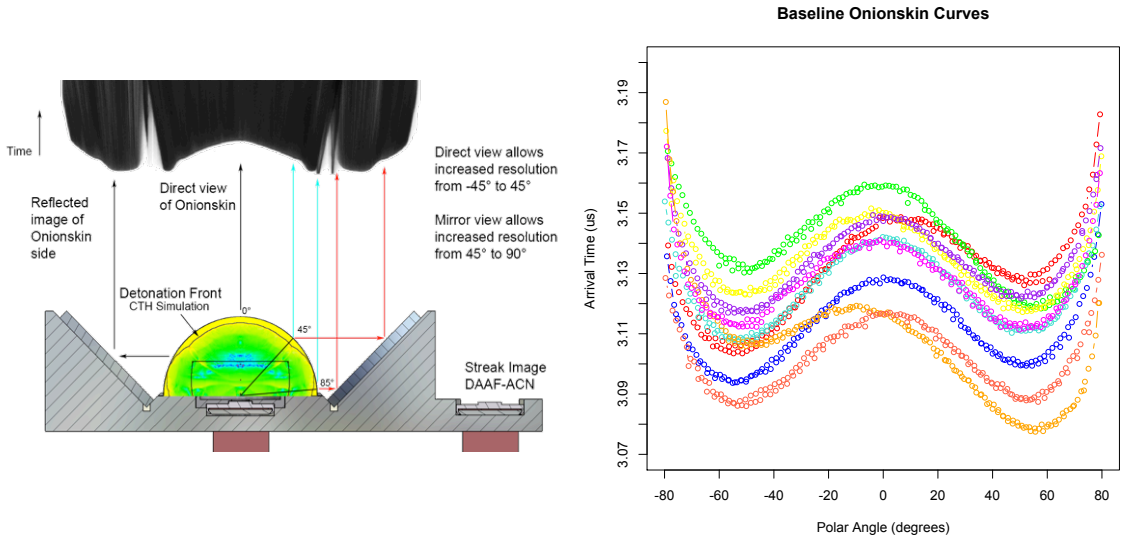
The problem to be addressed in this paper is the following. Approximately 30 onionskins, manufactured using very high quality materials, will be tested for the purpose of establishing a baseline. At various times in the future a new lot of onionskins will be manufactured using materials that are supposed to be of similar properties and quality to the baseline materials. A small sample of five to ten onionskins will be selected from the lot and tested. Before accepting the lot, engineers will need to determine if the onionskin curves from the test population are similar to those from the baseline population.

In this paper we develop and evaluate two approaches for testing the hypothesis that mean curves from two populations are the same. In addition, these two methods are extended to construct confidence bands for the mean curves and the difference in mean curves.

Figure 1: The Onionskin Test Setup. (a) The chamber used to carry out the explosions to collect onionskin curves (b) A 3D schematic of the onionskin unit sitting in the test device (c) A side view of the onionskin in a test device prior to testing and an image from a streak camera. (d) Onionskin data from historical tests.



(a) (b)



(c) (d)

The first approach is derived from more classical and pragmatic ideas. The standard approach in functional data analysis (FDA) (Ramsay & Silverman 2005) to compare two independent samples of noisy curves is to view each curve i as a p_i dimensional vector and register the arrival times of each curve at a common set of p angles, for example with cubic smoothing splines. In multivariate analysis, it is common to test for a mean difference with Hotelling's two-sample T^2 statistic (Christensen 2001). Of course, calculating T^2 requires

Table 1: Baseline data summaries.

Observation ID	Sample Size	Min Angle	Max Angle
15-2547	162	-83.0596	81.1393
15-2548	159	-83.8886	80.6162
15-2549	168	-86.6590	87.9967
15-2550	170	-83.8167	84.4541
15-2552	168	-83.3052	85.0207
15-2553	166	-82.8192	84.7257
15-2608	167	-85.9260	85.3868
15-2610	178	-88.5297	89.4079
15-2611	158	-82.9673	85.5605

the inversion of a $p \times p$ sample covariance matrix and generally p will be much larger than the number of baseline and lot sample curves. Therefore some kind of dimension reduction is usually required. Thus, the first approach models each curve as a linear combination of B-spline basis functions and then uses the T^2 statistic to compare the coefficients of the B-splines. The estimated coefficients also serve as a means to generate confidence bands for the mean curve. The first approach uses ideas very similar to those proposed in Besse, Cardot & Ferraty (1997) and Rice & Wu (2001), and its strength lies in its transparency and ease of implementation.

The second approach models each curve as a realization from a Gaussian Process (GP) (Stein 1999, Rasmussen & Williams 2006). This is similar in spirit to other Bayesian FDA approaches in the literature which treat the mean curve with a flexible model. For example Bigelow & Dunson (2007) model the basis coefficients in a multivariate adaptive spline model as normally distributed. Botts & Daniels (2008) proposed a similar knot selection type approach using B-splines. Morris & Carroll (2006) propose a wavelet-based random effects model, placing a normal distribution on the random wavelet coefficients. Our approach is the most similar to Behseta, Kass & Wallstrom (2005) who also avoid choosing an explicit set of basis functions and/or knot selection through use of a hierarchical GP. However, our model is different in a few important ways. We use a covariance function that allows an

explicit orthogonalization of the intercept and the shape of the functions. It is particularly critical in this problem that the overall level (i.e., intercept) of the mean curve be removed from consideration as will be discussed further in Section 2. This covariance function also lends itself well to a very efficient Gibbs sampling scheme which avoids the need to solve a large matrix at all during the MCMC iterations. Further, we propose a novel use of the Bayes Factor in a permutation test to compare two populations of curves, which is made feasible due to the efficient MCMC algorithm. The strength of this GP approach lies in its flexibility. Unlike the first approach it does not require the specification of a set of basis functions. However, the price paid is less transparency and additional computation time. Since the two approaches are complementary it is beneficial to apply them both to the onionskin problem, and ensure that the conclusions that can be drawn are the same.

The rest of the article is organized as follows. In Sections 2 and 3 the B-spline and GP approaches to comparing populations of curves and constructing confidence bands are presented, respectively. In Section 4 results from a simulation study to compare performances of the two approaches are presented. An actual analysis of onionskin populations is provided in Section 5, and Section 6 concludes the paper.

2 B-spline Basis Model for Onionskin Curves

In this section, a linear model using B-spline basis functions is described to represent each onionskin observation curve. For ease of presentation, only the model for the baseline observations is presented; the model for the test observations is identical. The i'th baseline observation is represented as

$$y_{i,j} = \beta_{i,0} + \sum_{k=1}^{5} \beta_{i,k} \mathcal{B}_k(\phi_{i,j}) + e_{i,j},\ i = 1, \ldots, m,\ j = 1, \ldots, n_i, \tag{1}$$

where $y_{i,j}$ is the arrival time for the i'th observation at the j'th observed angle $\phi_{i,j}$, $\mathcal{B}_k$ is the k'th B-spline basis function (defined below), and

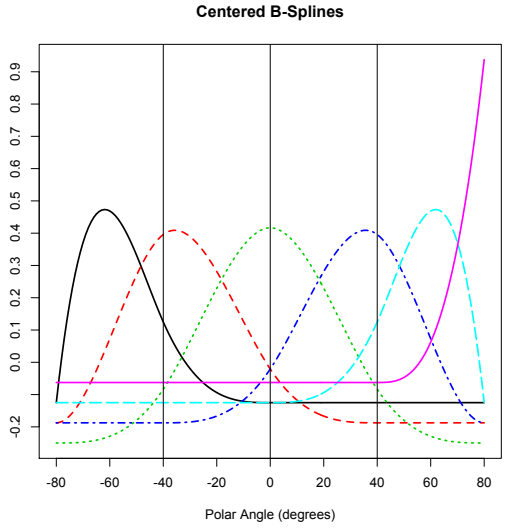
$$e_{i,j} \overset{iid}{\sim} N(0, \sigma^2)$$

is the measurement error of the arrival time. The $\beta_{i,k}$ are unknown regression coefficients that are estimated by ordinary least squares. Each observation has arrival times measured at different angles $\phi_{i,j}$, so the design matrix will be different for each observed curve. However, the basis *functions* are the same for all of the curves (baseline or test sample).

The actual basis used here is the B-spline basis (pp. 160-163, Hastie & Tibshirani (1990)), defined between $\pm 80°$ with interior knot locations located at $-40°, 0°$, and $40°$, and boundary knots at $\pm 80°$, resulting in a total of seven basis functions (six for the "shape" and one constant function for the intercept). However, the B-splines have also been adjusted by a constant shift so that they are orthogonal to constant functions in the L_2 function sense, i.e., all $\mathcal{B}_k$ have been vertically shifted so that

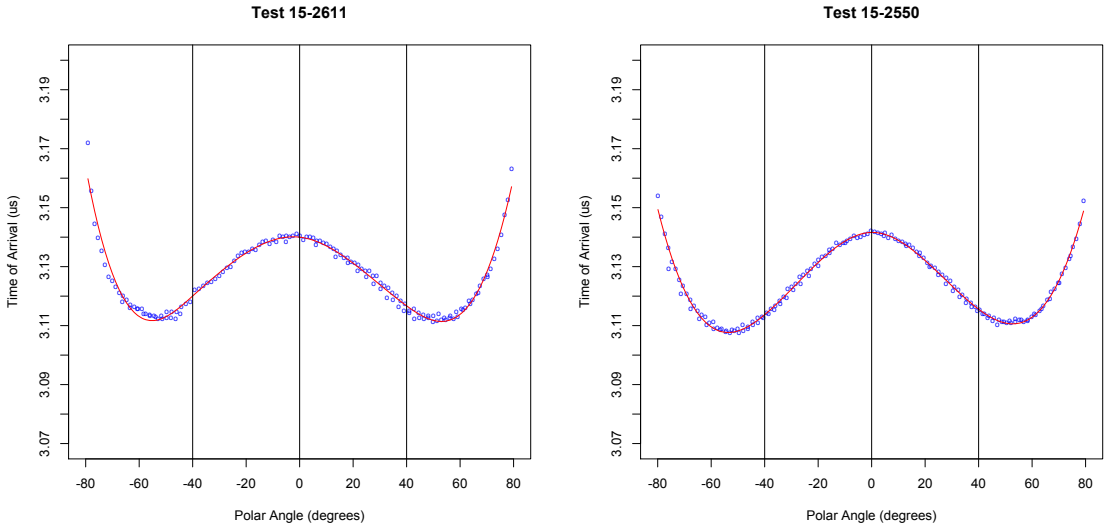
$$\int_{-80}^{80} \mathcal{B}_k(\phi) d\phi = 0.$$

Figure 2: Adjusted B-Spline basis functions, excluding the intercept. The vertical lines indicate the location of the interior knots.



It is important to use splines that are orthogonal to the intercept in this case. Ramsay & Silverman (2005), for example, recommend first aligning functions vertically when attempting to describe variability among functions. It is particularly critical to address this issue here, because in this application it is perfectly acceptable for the mean baseline coefficient $\alpha_0 = \mathrm{E}[\beta_{i,0}]$ to be different from the corresponding parameter in the test population $\alpha_0^* = \mathrm{E}[\beta_{i,0}^*]$, but they must have the same (or very similar) $\boldsymbol{\alpha}_s = \mathrm{E}[\beta_{i,1}, \ldots, \beta_{i,6}]^T$. That is, in this particular problem the curve populations must have similar shapes, irrespective of the overall level of the curves. The orthogonalization allows an unambiguous

Figure 3: Two examples of fitting a linear model with adjusted B-Splines. (a) Of the nine baseline curves the poorest fitting linear model is for test 15-2611 and (b) the best fitting model is for test 15-2550.



(a) (b)

separation of the "overall level" and "shape" effects.

Figure 2 shows a set of basis functions evaluated on a grid from $-80°$ to $80°$ in one degree increments. The constant shift adjustment is clearly seen in the plot. Figure 3 shows the data and fitted curves for two observations from the baseline population. These two examples the are "worst" and "best" fitting cases, respectively, as judged qualitatively from the baseline data. Fitted models for the other baseline curves are similar. The regression model with seven basis functions does an adequate job of fitting the onionskin data from the baseline population.

2.1 Comparing Spline Regression Coefficients

Hotelling's T^2 (Christensen 2001, Johnson & Wichern 2007) can be used to formally test if the mean vector of the baseline coefficients $\boldsymbol{\alpha}_s = \mathrm{E}[\boldsymbol{\beta}_i]$ is equal to the mean from a set of new onionskin tests $\boldsymbol{\alpha}_s^*$, where $\boldsymbol{\beta}_i = [\beta_{i,1}, \ldots, \beta_{i,6}]^T$ is the vector of shape coefficients for the i'th baseline observation (and $\boldsymbol{\beta}_i^*$ defined analogously). In this context Hotelling's T^2 is

$$T^2 = \frac{(m\, m^*)}{(m + m^*)} \, (\bar{\boldsymbol{\alpha}}_s \, - \, \bar{\boldsymbol{\alpha}}_s^*)^T \, \boldsymbol{S}^{-1} \, (\bar{\boldsymbol{\alpha}}_s \, - \, \bar{\boldsymbol{\alpha}}_s^*) \tag{2}$$

where (i) m and m^* are the baseline and test sample sizes, respectively, and (ii) $\bar{\boldsymbol{\alpha}}_s$ and $\bar{\boldsymbol{\alpha}}_s^*$ are the sample mean vectors of the estimated regression coefficients $\hat{\boldsymbol{\beta}}_i$ and $\hat{\boldsymbol{\beta}}_i^*$ for the baseline and test data, respectively, and (iii) $\boldsymbol{S}^{-1}$ is the inverse of the pooled sample covariance matrix of the estimated coefficients. With six basis functions (excluding the intercept) $\boldsymbol{S}$ is a 6×6 matrix and the inverse will exist with high probability if the number of curves (baseline or test) is at least seven (Okamoto 1973, Christensen 2001).

A subtle issue is that the $\boldsymbol{\beta}_i$ are not actually observed, rather their estimates $\hat{\boldsymbol{\beta}}_i$ are. However, from standard linear model theory $\mathrm{E}[\hat{\boldsymbol{\beta}}_i] = \mathrm{E}\{\mathrm{E}[\hat{\boldsymbol{\beta}}_i \mid \boldsymbol{\beta}_i]\} = \mathrm{E}\{\boldsymbol{\beta}_i\} = \boldsymbol{\alpha}_s$ so that inference about $\boldsymbol{\alpha}_s$ can be carried out using the $\hat{\boldsymbol{\beta}}_i$. If the $\hat{\boldsymbol{\beta}}_i$, (and $\hat{\boldsymbol{\beta}}_i^*$) are multivariate normal, and if $\boldsymbol{\alpha}_s = \boldsymbol{\alpha}_s^*$, then a multiple of T^2 will have a central F distribution (Christensen 2001). However, because the sample sizes are too small to adequately assess multivariate normality, we also consider a permutation test (Mielke & Berry 2001) for $H_0 : \boldsymbol{\alpha}_s = \boldsymbol{\alpha}_s^*$, using the T^2 test statistic. This proceeds as follows.

Consider all $\binom{C=m+m^*}{m}$ combinations $c = 1, \ldots, C$ of grouping the $m + m^*$ observations into two sets of m and m^* observations, respectively. For each combination of observations c calculate T^2 as defined in (2), and denote this $T^2(c)$. The permutation p-value is defined as

$$p = \frac{1}{C} \sum_{c=1}^{C} I\{T^2(c) \geq T^2\}, \tag{3}$$

where $I(A)$ is the indicator of the set A ($I(A) = 1$ if A occurs and 0 otherwise), and T^2 is the value of the test statistic on the original sample. For computational efficiency, it is common to take a smaller random sample of all C combinations and approximate the p-value based on this subset of all possible combinations. Because computation of $T^2(c)$ for a given permutation is quite fast, p-values are calculated based on 10,000 randomly selected permutations for the presented results.

The permutation test is guaranteed to keep a specified type I error rate, under only the assumption of exchangeability. That is, the null hypothesis is technically that the baseline and test populations have the same distribution (i.e., not necessarily normal). However, the T^2 statistic is explicitly formulated to test for changes in $\boldsymbol{\alpha}_s$, so that this test will have power to detect the difference of primary concern, namely a difference in mean "shape". The normal theory p-value approach is compared to the permutation test p-value approach (and the GP approach described next in Section 3) in a simulation study provided in Section 4.

2.2 Simultaneous Confidence Bands for Mean Curves

A procedure for constructing $100\,(1-\gamma)\,\%$ confidence bands for the baseline mean "shape" curve is described below. The procedure for the test data is entirely analogous. The construction of confidence bands for the difference in mean shape curves is also discussed.

Let $\bar{\alpha}_{s,j}$ and s_j be the sample average and standard deviation, respectively, of the estimates for the j'th coefficient for each observation, $\hat{\beta}_{1,j}, \ldots, \hat{\beta}_{m,j}$. With six basis functions $100\,(1-\gamma)\,\%$ Bonferroni confidence intervals for the mean coefficients $\alpha_{s,j}$ are

$$\bar{\alpha}_{s,j} \pm t(1-(\gamma/6), m-1)\, s_j/\sqrt{m}, \quad j = 1, \ldots, 6 \tag{4}$$

where $t(r, \nu)$ indicates the r'th quantile of a t distribution with ν degrees of freedom. Taken together the intervals define a $100\,(1-\gamma)\,\%$ Bonferroni confidence region (i.e., hypercube) for the mean coefficient vector $\boldsymbol{\alpha}_s$. Next, sample N coefficient vectors from the Bonferroni region and generate N mean curves with adjusted B-spline basis functions defined over a grid of 161 equally spaced angels from $-80°$ to $80°$. Upper and lower confidence curves are determined by finding the maximum and minimum of the N curves at each grid angle. The resulting confidence curves are *conservative* simultaneous confidence bounds in the sense that *at least* $100\,(1-\gamma)\,\%$ of them will contain the true mean curve under repeated sampling (and under all of the model assumptions). The resulting bounds are approximate (due to the brute force sampling step), but the approximation error can be made arbitrarily small by

making N large. The following results used $N = 10^6$ which were negligibly different to results obtained using $N = 10^5$. Confidence Bands could also be constructed by instead sampling coefficients from the simultaneous T^2 ellipse, however, these bands are generally wider than those resulting from the Bonferroni hypercube for this problem. Confidence bands for the mean curve of the test population mean curve are found similarly.

To construct confidence bands for the difference in mean shape curves, calculate a Bonferroni confidence region for the vector of differences $\boldsymbol{\delta}_s = \boldsymbol{\alpha}_s - \boldsymbol{\alpha}_s^*$ and apply the same process as above (i.e., randomly generating N values of $\boldsymbol{\delta}_s$ uniformly from the Bonferroni hypercube and converting them to difference curves on a $\{-80, \ldots, 80\}$ angle grid). The upper and lower confidence bands for the mean shape difference are determined by finding the maximum and minimum of the N difference curves at each grid angle. Confidence bands for onionskin problem are presented in Section 5.

3 Gaussian Process Model for Onionskin Curves

In this section, a hierarchical GP model for the onionskin curves is presented. It is assumed that each observation is generated as a Gaussian process random effect curve plus noise off of a parent mean curve, which is also modeled as GP. This approach has the advantage of not having to select a basis representation for the curves as in the previous section. The price paid is additional computation time, which is discussed further in Section 4. Methods for comparing the model assuming different shaped mean curves for baseline and test populations to the model assuming the same shaped mean curves are discussed in Section 3.2.

3.1 The Hierarchical GP Model for Onionskin Curves

Once again, for ease of presentation, the model for the baseline data is presented first; the model for the test data will be identical. The i'th onionskin observation (in the baseline data) is represented as

$$y_{i,j} = \mu(\phi_{i,j}) + \nu_i(\phi_{i,j}) + e_{i,j},\ i = 1, \ldots, m,\ j = 1, \ldots, n_i \tag{5}$$

where μ is a mean function shared among all $m = 9$ baseline observations, ν_i is the random effect deviation curve off of the mean curve for the i'th observation, and $e_{i,j}$ is *iid* measurement error.

The mean curve μ in (5) is modeled as

$$\mu(\phi) = \alpha_0 + \alpha(\phi), \tag{6}$$

where

$$\alpha_0 \sim N(\dot{a}, \tau_0^2), \tag{7}$$

$$\alpha_s \sim GP(\dot{b}, K_{\boldsymbol{\tau}}), \tag{8}$$

and $GP(a, K)$ denotes a Gaussian Process with mean function a and covariance function K. Here and below, all prior parameters that will eventually be specified later in this section (e.g., $\dot{a}$ and $\dot{b}$ above) are denoted with a dot above them. The covariance function $K_{\boldsymbol{\tau}}$ is

$$K_{\boldsymbol{\tau}}(\phi, \phi') = \left[\tau_1^2 \mathcal{P}_1(\phi)\mathcal{P}_1(\phi') + \tau_2^2 \mathcal{P}_2(\phi)\mathcal{P}_2(\phi')\right] - \tau_3^2 \frac{\mathcal{P}_4(|\phi - \phi'|)}{4!}, \tag{9}$$

where $\mathcal{P}_k$ is the k'th Bernoulli polynomial, and

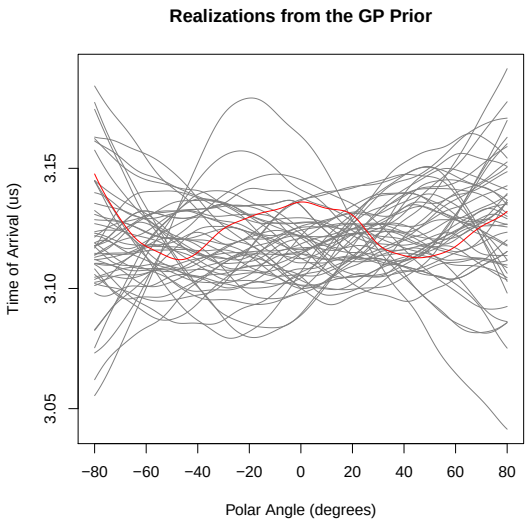
$$\tau_j^2 \overset{ind}{\sim} \text{InvGamma}(\dot{c}_j, \dot{d}_j),\ j = 0, 1, 2, 3. \tag{10}$$

The covariance function in (9) is the same as that used in Reich, Storlie & Bondell (2009) for Bayesian smoothing spline (BSS) ANOVA models and also used in Storlie, Michalak, Quinn, DuBois, Wender & DuBois (2012). The motivation for this form of the covariance comes from the relation of GP regression to smoothing splines (Wahba 1990). Steinberg & Bursztyn (2004) also demonstrate that this Gaussian process model is equivalent to a Bayesian trigonometric regression model with independent Gaussian priors for the trigonometric basis

functions' coefficients with variances that decrease as the frequencies of the trigonometric functions increase.

Figure 4 provides some conceptual realizations from the the GP prior for μ. The curves in the figure were *not* actually generated from the unconditional prior used in this study since the diffuse prior actually used for α_0 in (7) and the τ_j (10) would not provide the appropriate context. Hence, the curves in the plot are the result of randomly drawing α_s conditional on fixing α_0 and the τ_j to reasonable values (i.e., their resulting posterior mean using the proposed Hierarchical GP model given the data in Figure 2(d)). Thus, the curves in the figure give an idea of the type of α_s curves that *can* be produced from the prior if given feasible values for the hyper prior parameters.

Figure 4: Conceptual realizations from the GP "prior". The curve most qualitatively representative of the onionskin curves among the 50 realizations in the plot is highlighted in red for illustration.



There are three major advantages to using the BSS-ANOVA covariance over the traditional powered exponential covariance (or similar) in this case, (i) it is much more computationally stable when domain points ($\phi_{i,j}$) are close together as is the case here, (ii) it lends itself to a much faster MCMC algorithm as discussed further in Appendix A, and (iii) its support consists of functions α_s that are orthogonal in the L_2 function sense to constant functions (i.e., every draw from this α_s's prior and hence posterior is orthogonal to the intercept term α_0). The third advantage is important in this case, as mentioned above, since again it is perfectly acceptable for the baseline α_0 to be different from the corresponding parameter in the test population α_0^*, but they must have the same (or very similar) α_s curves. The orthogonal separation of intercept and shape makes the comparison of baseline and test populations' shapes straightforward and unambiguous. For example, if $\alpha_s \neq \alpha_s^*$, then it is

certain that baseline and test populations have different shapes. On the other hand, if the α_s were not orthogonal to constant functions, then $\boldsymbol{\alpha}_s \neq \boldsymbol{\alpha}_s^*$ could possibly mean that the shapes are actually the same, and the mean curves only differ by an additive constant.

The model used for the random effect curves ν_i is entirely similar to that for μ, namely

$$\nu_i(\phi) = \beta_{0,i} + \beta_{s,i}(\phi), \tag{11}$$

where

$$\beta_{0,i} \overset{iid}{\sim} N(0, \varsigma_0^2), \tag{12}$$

$$\beta_{s,i} \overset{iid}{\sim} GP(0, K_{\boldsymbol{\varsigma}}), \tag{13}$$

with $K_{\boldsymbol{\varsigma}}$ defined for $\boldsymbol{\varsigma} = [\varsigma_1, \varsigma_2, \varsigma_3]'$ by simply replacing each τ_j with ς_j in (9), and

$$\varsigma_j^2 \overset{ind}{\sim} \text{InvGamma}(\dot{e}_j, \dot{f}_j),\ j = 0, 1, 2, 3. \tag{14}$$

Finally, for the measurement error it is assumed that

$$e_{i,j} \overset{iid}{\sim} N(0, \sigma^2) \tag{15}$$

$$\sigma^2 \sim \text{InvGamma}(\dot{g}, \dot{h}). \tag{16}$$

To complete the model specification for $y_{i,j}$ prior parameter values $\dot{a}$, $\dot{b}$, $\dot{c}_j$, $\dot{d}_j$, $\dot{e}_j$, $\dot{f}_j$, $j = 0, 1, 2, 3$, $\dot{g}$, and $\dot{h}$ must be specified. In this paper, vague but proper priors are used for all of the parameters, since there was little previous data or information available about these onionskin curves. The priors for α_0 and α_s in (7) and (8), were set with $\dot{a} = 0$ and $\dot{b} \equiv 0$, respectively. For the variance parameters τ_j, $j = 0, 1, 2, 3$, in (10) the priors were set with $[\dot{c}_j, \dot{d}_j] = [1, 1]$ for each j. Analogously, the same vague priors were placed on the parameters related to ν_i, i.e., $[\dot{e}_j, \dot{f}_j] = [1, 1]$ in (14). Lastly, $[\dot{g}, \dot{h}] = [0.1, 0.1]$ were set to put a very vague prior on σ^2, since there is actually a lot of information about this parameter anyhow (e.g., each of the $m = 9$ baseline observations has ~ 160 data points that provide

information about σ^2).

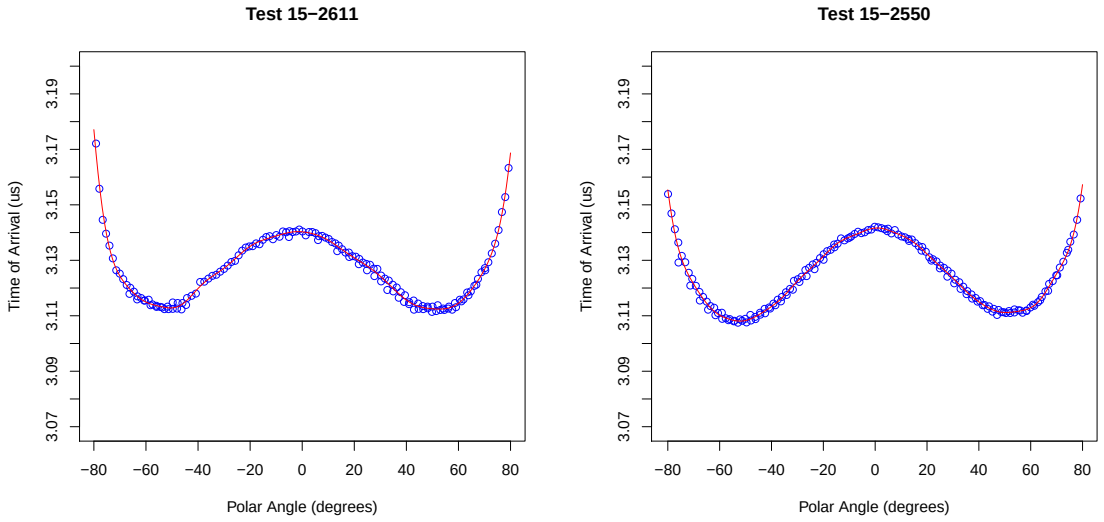
An efficient Gibbs sampling algorithm (see Givens & Hoeting (2000), for example) to obtain an approximation to the posterior distribution of the parameters of the model above is described in Appendix A. This sampling approach deviates from the standard MCMC approach for GPs (such as that described in Behseta et al. (2005)) which would discretize the domain into a grid and sample from a large multivariate normal. In the hierarchical model proposed here this would be very computationally expensive, since in order to MCMC sample from the mean curve, it would require conditioning on the current curve values at all grid points from all m curves for each unit. Instead, we take advantage of the structure of the covariance in (9), and ultimately sample the coefficients from a truncated Karhunen-Loéve expansion (Berlinet & Thomas-Agnan (2004), pp.65-70). Thus, the flexible GP approach described above can be reduced to a simple random effects linear model for computational purposes, for which a Gibbs sampling scheme can be easily implemented (Zeger & Karim 1991). For practical purposes, this is in some ways the same as just starting from an over-parameterized flexible linear model (e.g., spline basis) and treating the coefficients as random effects, using the covariance of the coefficients to encourage smoothness, for instance. However, in starting from the GP, the appropriate covariance to give the basis coefficients is simply dictated by the GP covariance, and this covariance is also diagonal which increases computational efficiency. More details are provided in Appendix A.

Figure 5 shows the data and the posterior mean for $\mu + \nu_i$ from the GP approach for the same two baseline curves displayed in Figure 2(d). It is apparent from the plots that the GP model is general enough to represent the onionskin curves very well.

For completeness, the i'th observation from the *test* population is represented as

$$y^*_{i,j} = \mu^*(\phi_{i,j}) + \nu^*_i(\phi_{i,j}) + e^*_{i,j},\ i = 1, \ldots, m^*,\ j = 1, \ldots, n^*_i, \tag{17}$$

Figure 5: Two examples of fitting the Hierarchical GP model to the baseline curves. For consistency with Figure 3 the curves from test 15-2611 and test 15-2550 are presented.



(a) (b)

with

$$\mu^*(\phi) = \alpha_0^* + \alpha_s^*(\phi), \tag{18}$$

and all priors for α_0^* and $\alpha_s^*(\phi)$ *iid* to those for α_0 and $\alpha_s(\phi)$, respectively.

3.2 Comparing Baseline and Test Populations

As in Section 2.1, a test is described for whether or not the shape of the mean curve for the baseline population equals that for the test population, i.e., $H_0 : \alpha_s \equiv \alpha_s^*$ vs. $H_A : \alpha_s \neq \alpha_s^*$. It is common in Bayesian Statistics to use the Bayes Factor (BF) (Kass & Raftery 1995) to compare one hypothesis (model) to another,

$$BF = \frac{p(\mathcal{Y} \mid \alpha_s \neq \alpha_s^*)}{p(\mathcal{Y} \mid \alpha_s = \alpha_s^*)} \tag{19}$$

where $\mathcal{Y}$ is a set of all of the baseline and test data. If $BF > 1$, it provides evidence for H_A, and Kass & Raftery (1995) provide some guidelines for the strength of the evidence indicated by the value of BF. Computing BFs can be challenging when the parameter space is of large dimension. Kass & Raftery (1995) review several approximations to obtain BF which all

have some weaknesses, related to either accuracy or numerical instability. Carlin & Chib (1995) provide a Gibbs sampling algorithm, which samples an indicator for which model is correct at each MCMC iteration. The Carlin-Chib algorithm makes use of the MCMC approximated posterior under each model (H_0 and H_A), and the prior model probabilities can be adjusted to guarantee good mixing, thus giving a reliable estimate of BF even if it is very small or large. In Appendix A a Gibbs sampling algorithm for the approximation of the posterior distribution of the model parameters defined in Section 3.1 is described, and a straight-forward application of the Carlin-Chib algorithm can then be used to calculate BF.

Even though the Carlin-Chib algorithm alleviates the instability issues with BF calculation, it is also well known that BFs can be very sensitive to prior specifications (Sinharay & Stern 2002). This is particularly true in our case (as can be seen in the results of Section 4), since we were forced to use a vague prior. Because of the vague priors, small differences between α_s and α_s^* result in very small values for BF ($\sim 10^{-10}$). This is because μ and μ^* can look like just about anything under vague priors. Hence, when they are reasonably close to each other (even if they **are** different) the posterior probability favors the model that treats them as the same.

However, we noticed that though these values for BF were small, they were substantially larger for test cases where $\alpha_s \neq \alpha_s^*$ relative to cases where $\alpha_s \equiv \alpha_s^*$. This suggested the novel idea of using BF as a test statistic in a permutation test (Mielke & Berry 2001). That is, perform the same permutation test as that described in Section 2 only with T^2 replaced by BF. Because the computation of a BF takes much longer than that for a T^2 statistic, the sample of all possible combinations in this case is limited to 1000. A log-normal distribution is then fit to the sample of BFs resulting from each of the combinations $c = 1, \ldots, 1000$, and finally the permutation p-value is calculated according to

$$p = 1 - \Phi((\log(BF) - a)/b), \tag{20}$$

where Φ is the standard normal CDF and a and b are the mean and standard deviation, respectively, of the BFs for each combination ($BF(c)$, $c = 1, \ldots, 1000$). This procedure can be time consuming (more details provided in Section 4), but it is also trivially parallelizable, which can ease the computational burden substantially.

While this is essentially now a frequentist hypothesis test, it is based on a test statistic (BF) that accurately reflects our knowledge of the parameters in a posterior sense. It seems unfortunate that a lack of substantial a-priori information, should have to drive the type I error rate (and hence power for subtle differences) to essentially zero (e.g., if using a rule such as reject H_0 if $BF > 1$). The permutation test, on the other hand is guaranteed to have type I error rate equal to a pre-specified level. It also happens to have impressive power to detect differences between α_s and α_s^* as will be seen in Section 4. In future studies, we will be able to use the information gained in this study to provide more informative priors, in which case the BF may be more useful by itself.

If there is evidence to suggest that $\alpha_s \neq \alpha_s^*$, then it is helpful to further understand the nature of difference. To this end, a procedure is described for producing simultaneous Bayesian credible bands for α_s, α_s^*, and the difference between them $\delta_s = \alpha_s^* - \alpha_s$. The functions (α_s^l, α_s^u) are $(1-\gamma)100\%$ simultaneous credible bands for α_s if

$$\Pr(\alpha_s^l(t) < \alpha_s(t) < \alpha_s^u(t), \text{ for all } t \in T \mid \mathcal{Y}) = 1 - \gamma. \tag{21}$$

where T is the domain of the curves (i.e., $T = [-80, 80]$ in this case).

The MCMC algorithm described in Appendix A provides an approximate posterior sample of $\alpha_s(\boldsymbol{t})$, $\alpha_s^*(\boldsymbol{t})$, and $\delta_s(\boldsymbol{t})$, for a dense grid $\boldsymbol{t} = [t_1, \ldots, t_g]'$. Denote this posterior sample as $[\alpha_{s,1}(\boldsymbol{t}), \alpha_{s,1}^*(\boldsymbol{t}), \delta_{s,1}(\boldsymbol{t})], \ldots, [\alpha_{s,N}(\boldsymbol{t}), \alpha_{s,N}^*(\boldsymbol{t}), \delta_{s,N}(\boldsymbol{t})]$. Approximate $(1-\gamma)100\%$ simultaneous confidence bands for α_s can then be obtained by finding the largest value of γ' satisfying

$$\frac{1}{N}\sum_{n=1}^{N} I\{\xi_{l,\gamma'/2} < \alpha_{s,n}(t_l) < \xi_{l,1-\gamma'/2}, \text{ for each } l = 1, \ldots, g\} < 1 - \gamma \tag{22}$$

where $\xi_{l,p}$ is the p'th sample quantile of the approximate posterior distribution of $\alpha_s(t_l)$. Using the largest value of γ' satisfying (22) set

$$\alpha_s^l(t_l) = \xi_{l,\gamma'/2} \text{ for } l = 1, \ldots, g \tag{23}$$

$$\alpha_s^u(t_l) = \xi_{l,1-\gamma'/2} \text{ for } l = 1, \ldots, g, \tag{24}$$

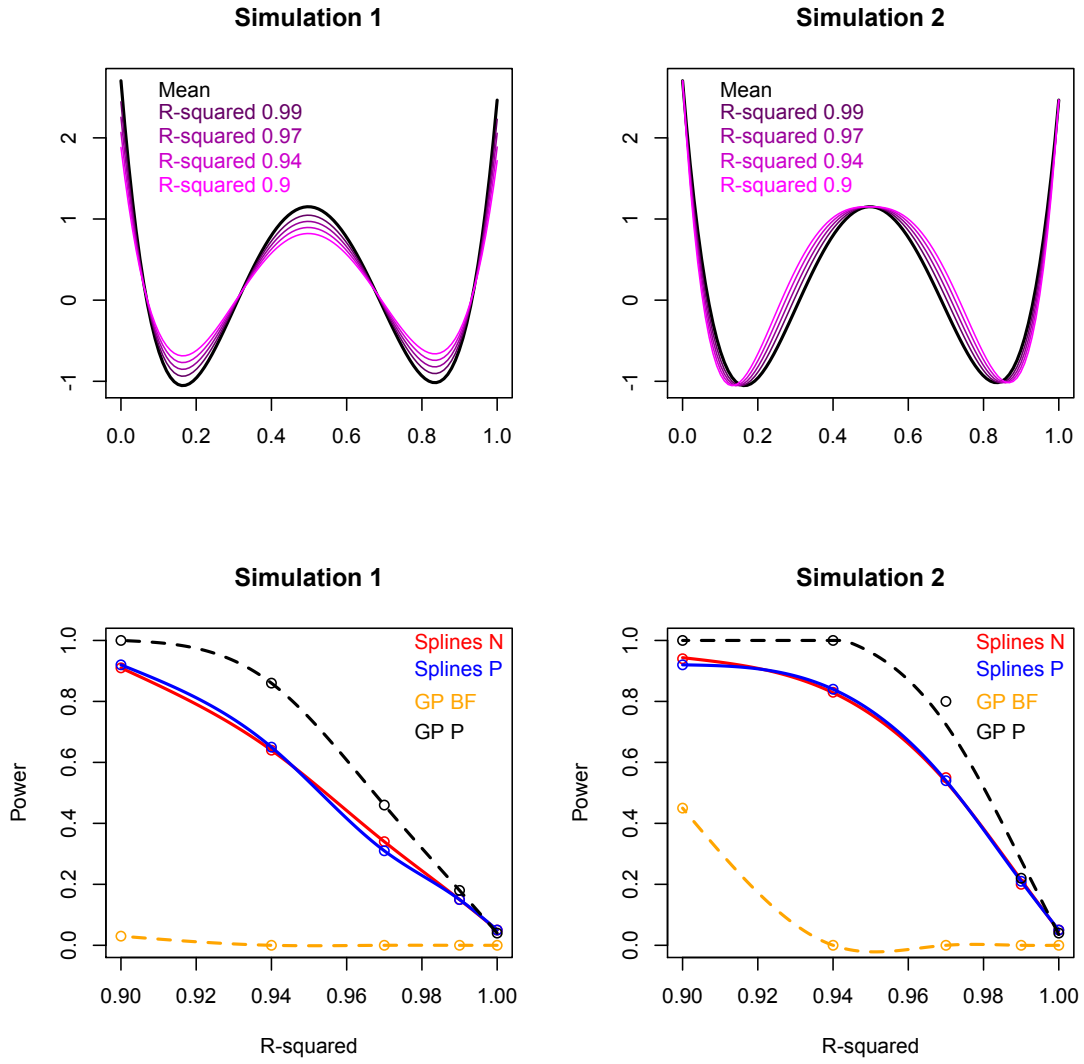
then $(\alpha_s^l(\boldsymbol{t}), \alpha_s^u(\boldsymbol{t}))$ provide approximate (due to MCMC sampling error, and a finite $\boldsymbol{t}$ grid) simultaneous credible bands for α_s. Simultaneous credible bands for α_s^* and δ_s can be constructed in the same manner.

4 Simulation Study

Two simulation studies were conducted to compare the B-spline approach to the GP approach. The hypothesis being tested is that the baseline and test data mean curves have similar shape. For the B-spline approach, the observed T^2 is compared to the normal theory critical value and the permutation test critical value. For the GP method, BF is computed and H_0 rejected if $BF > 1$, and also the BF is compared to the permutation test critical value. The number of basis functions in the MCMC approximation discussed in Appendix A was set to $N = 50$, and the total number of MCMC iterations was fixed at 10,000. In both studies the baseline mean curve was constructed to be similar to the baseline mean curve of the data presented above. Figure 6 shows the baseline mean curve and various test data mean curves for the two scenarios and the resulting power curves.

In both studies the squared "correlation" or R^2 between arrival times (i.e., assuming a uniform distribution on angle) of the test data mean curve to the baseline mean curve varied from 0.9 to 1.0 (i.e., the null case). In the first study the mean curves were adjusted to differ at the pole (0.5 in the plot) and also near the minima. In the second study the mean curves had the same value at the pole and the minima, but differed in-between.

Figure 6: Representative curves from the simulation studies and power curves shown as a function of R^2. In the legend "N" denotes a power curve based on normal theory; a "P" indicates the power is based on a permutation test and "BF" is the power from the Bayes factor approach.



The GP with a permutation test (GP-P) was the most powerful test in both studies. As would be expected under normality of the βs, the normal theory and permutation test approaches using B-splines (Splines-N and Splines-P, respectively) give essentially the same results. The GP-BF approach has almost no power to detect a difference in mean curves except in study two when the correlation between curves is 0.9. As mentioned previously, this is an artifact of the vague prior placed on the parameters of the GP model. If more informative priors are used (e.g., in a future study) the GP-BF method may be more useful.

In general there is more power to detect differences in mean curves in the second scenario than in the first. In either case, however, the tests (with the exception of GP-BF) are able to effectively identify very subtle differences between baseline and test populations.

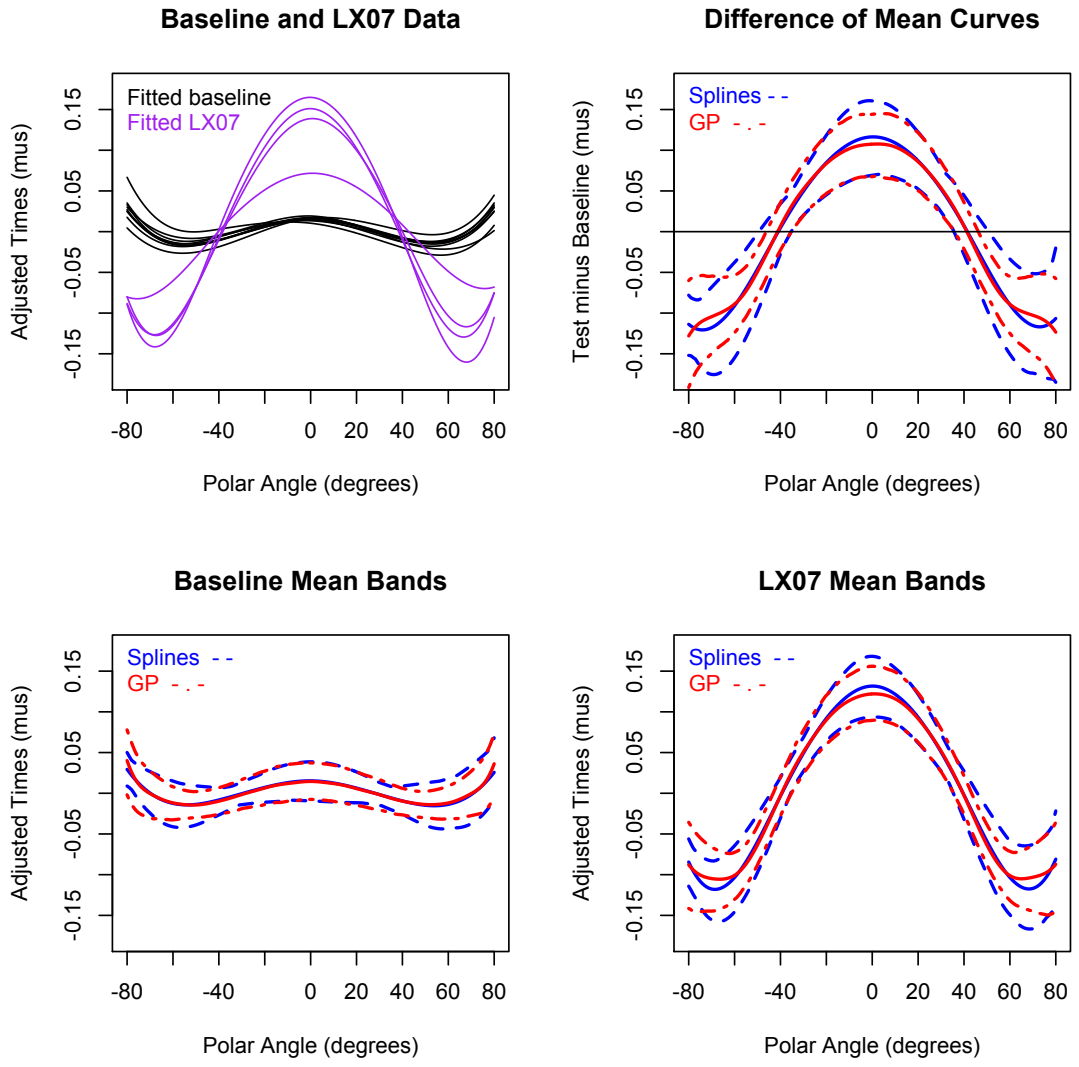
While the GP-P test was the best performer in terms of power, it was substantially more time consuming than the other methods to conduct the hypothesis test. The computation times to compute the hypothesis test for a single realized dataset from the first simulation for methods Splines-N, Splines-P, GP-BF and GP-P, were $\sim$0 minutes, 2 minutes, 3 minutes, and 65 minutes, respectively. However, the reported GP-P time is from a parallel run on a 48-way commodity machine, so that its total CPU time used was actually 3120 minutes.

5 Illustration: Comparing Baseline and New Data

In this section, the baseline population is compared to a population of onionskin tests that used a different detonator (an LX07 detonator). Because the two sets of onionskin tests used different detonators, the mean curves are expected to have differences. Confidence bands for each population mean curve and difference between means are also constructed using the B-spline and GP approaches.

The p-values from the three tests, Spines-N, Splines-P and GP-P, were all $\ll 0.001$, a strong indication of a mean shape difference between the two populations. Figure 7 shows the fitted baseline and LX07 curves excluding the intercept (top left plot). The top right plot shows confidence bands for the difference in mean curves. If the two mean shape curves were the same, the confidence bands would be expected to contain a horizontal line at zero. Clearly, the LX07 mean curve is different from the baseline mean curve. The two bottom plots show confidence bands for the individual mean curves. The spline and GP models produce confidence bands that are very similar, providing some additional assurance that the conclusions drawn are accurate.

Figure 7: Confidence bands for baseline and LX07 test data. The top left panel shows the fitted curves (posterior means from the GP method), excluding the intercept. The top right panel shows 95% confidence bands for the difference in mean shape curves. The two bottom panels show 95% confidence bands for the mean shape curves.



6 Summary and Conclusions

In this article we have developed and compared two methods for estimating and comparing the shape of two mean curves from possibly different populations based on independent samples from the populations. Both approaches were then used to analyze two populations of onionskins. The B-spline method models each sample curve as a linear combination of

basis functions and then uses the estimated regression coefficients as the basis for hypothesis testing and construction of simultaneous confidence bands. The second approach models each sample curve as a more general hierarchical GP, uses the Bayes Factor (by itself and in a permutation test) for comparing populations, and uses the approximate posterior distribution obtained via Gibbs sampling to obtain simultaneous credible Bands. Both approaches produced very similar "confidence" bands for the mean curves and for the difference in means curves for the onionskin analysis. Empirical evidence from the simulation study suggests that the novel GP approach based on a permutation test is more powerful for detecting differences in means curves than the B-spline approach. This increase in power, however, comes at the expense of much higher computational cost.

While the main focus of this paper has been to compare the shape of mean curves from possibly different populations it is also possible that interest lies in detecting differences in variability. For example, we might ask if the variability seen in a sample of observation curves from a given lot is the same as in baseline observation curves? Using the B-spline approach, the covariance matrices of the estimated regression coefficients, including the intercept, could be tested for equality between the baseline and test populations. This test could be easily combined with the test for difference in mean curves proposed here, if a single test is desired, using a simple Bonferroni correction. In the GP approach, incorporating a test for difference in variance parameters of curves across populations into the Bayes Factor could easily be accomplished by fitting the restricted model assuming the same variance parameters as well as mean shape parameters.

Acknowledgments

The authors' work was funded by the C-8 Enhanced Surveillance program of Los Alamos National Security, LLC (LANS), operator of the Los Alamos National Laboratory under Contract No. DE-AC52-06NA25396 with the U.S. Department of Energy. This paper is

published under LA-UR-11-05417.

References

Behseta, S., Kass, R. E. & Wallstrom, G. L. (2005), 'Hierarchical models for assessing variability among functions', *Biometrika* **92**(2), 419–434.

Berlinet, A. & Thomas-Agnan, C. (2004), *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Kluwer Academic Publishers, Norwell, MA.

Besse, P. C., Cardot, H. & Ferraty, F. (1997), 'Simultaneous non-parametric regressions of unbalanced longitudinal data', *Computational Statistics & Data Analysis* **24**(3), 255 – 270.

Bigelow, J. L. & Dunson, D. B. (2007), 'Bayesian adaptive regression splines for hierarchical data', *Biometrics* **63**(3), 724–732.

Botts, C. H. & Daniels, M. J. (2008), 'A flexible approach to bayesian multiple curve fitting', *Computational Statistics & Data Analysis* **52**(12), 5100 – 5120.

Carlin, B. P. & Chib, S. (1995), 'Bayesian model choice via markov chain monte carlo methods', *Journal of the Royal Statistical Society. Series B (Methodological)* **57**(3), pp. 473–484.

Christensen, R. (2001), *Advanced Linear Modeling*, 2nd edn, Springer-Verlag, New York, NY.

Dallman, J. (1988), 'Measurements of detonation-wave spreading and local particle velocity at the surface o f i7-mm lx-07 hemispherical boosters', *Los Alamos National Laboratory Technical Report, LA-11414-MS* .

Givens, G. & Hoeting, J. (2000), *Computational Statistics*, 1st edn, Hoboken, NJ: John Wiley & Sons.

Hastie, T. & Tibshirani, R. (1990), *Generalized Additive Models*, Chapman & Hall/CRC.

Hogan, G., Adams, K. & Alrick, K., et. al. (1999), 'Proton radiography', *Proceedings of the 1999 Particle Accelerator Conference* **1**, 579–583.

Hooks, D. E., Morris, J. S., Hill, L. G. & Francois, E. (2008), 'Performance evaluation of booster materials in the plastic bonded explosive pbx 9502 in a hemispherical wave breakout test', *Propellants, Explosives, and Pyrotechnics* pp. 1–14.

Johnson, R. & Wichern, D. (2007), *Applied Multivariate Statistical Analysis*, 6th edn, Upper Saddle River, NJ: Pearson Prentice Hall.

Kass, R. E. & Raftery, A. E. (1995), 'Bayes factors', *Journal of the American Statistical Association* **90**(430), pp. 773–795.

Lundberg, A. (1996), 'High explosives in stockpile surveillance indicate constancy', *Science & Technology Review, Lawrence Livermore National Laboratory* **12**, 12.

Mielke, P. & Berry, K. (2001), *Permutation Methods*, New York, NY: Springer-Verlag.

Morris, J. S. & Carroll, R. J. (2006), 'Wavelet-based functional mixed models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(2), 179–199.

Okamoto, M. (1973), 'Distinctness of the eigenvalues of a quadratic form in a multivariate sample', *Annals of Statistics* **1**, 763–765.

Ramsay, J. & Silverman, B. (2005), *Functional Data Analysis*, New York, NY: Springer-Verlag.

Rasmussen, C. & Williams, C. (2006), *Gaussian Processes for Machine Learning*, MIT Press.

Reich, B., Storlie, C. & Bondell, H. (2009), 'Variable selection in Bayesian smoothing spline ANOVA models: Application to deterministic computer codes', *Technometrics* **51**, 110–120.

Rice, J. A. & Wu, C. O. (2001), 'Nonparametric mixed effects models for unequally sampled noisy curves', *Biometrics* **57**(1), 253–259.

Sinharay, S. & Stern, H. S. (2002), 'On the sensitivity of bayes factors to the prior distributions', *The American Statistician* **56**, 196–201.

Stein, M. (1999), *Interpolation of Spatial Data*, Springer-Verlag, New York, NY.

Steinberg, D. & Bursztyn, D. (2004), 'Data analytic tools for understanding random field regression models', *Technometrics* **46**, 411–420.

Storlie, C. B., Michalak, S. E., Quinn, H. M., DuBois, A. J., Wender, S. A. & DuBois, D. H. (2012), 'A bayesian reliability analysis of neutron induced errors in high performance computing hardware', *Journal of the American Statistical Association (in review)* .

Wahba, G. (1990), *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series in Applied Mathematics.

Zeger, S. L. & Karim, M. R. (1991), 'Generalized linear models with random effects; a gibbs sampling approach', *Journal of the American Statistical Association* **86**(413), pp. 79–86.

A Supplementary Material

A.1 Sampling from the Hierarchical GP Model

The Karhunen-Loéve Theorem (Berlinet & Thomas-Agnan (2004), pp. 65-70) guarantees that any Gaussian process $X(t)$, $t \in [a, b]$ with mean function $\mu(t)$ continuous covariance function $K(s, t)$ can be represented as

$$X(t) = \mu(t) + \sum_{k=1}^{\infty} Z_k \psi(t), \tag{25}$$

where the (i) $Z_k \overset{ind}{\sim} N(0, \lambda_k)$, (ii) λ_k and ψ_k are the eigenvalues and eigenfunctions of K, respectively, $k = 1, 2, \ldots$. The expression in (25) suggests the approximation

$$X(t) \approx \mu(t) + \sum_{k=1}^{N} Z_k \psi(t), \tag{26}$$

for some N. It is not imperative that N be extremely large in practice, since it need only be large enough to allow enough high frequency eigenfunctions to represent the curve suggested by the data.

The form of the covariance for the proposed GP model in (9) is additive, i.e., it can be written as

$$K_{\boldsymbol{\tau}}(s, t) = \tau_1^2 K_1(s, t) + \tau_2^2 K_2(s, t) + \tau_3^2 K_3(s, t), \tag{27}$$

with $K_1(s, t) = \mathcal{P}_1(s)\mathcal{P}_1(t)$, $K_2(s, t) = \mathcal{P}_2(s)\mathcal{P}_2(t)$, and $K_3(s, t) = -\mathcal{P}_4(|s - t|)/4!$. Recall that the mean curve in the hierarchical model in (5) is $\mu(t) = \alpha_0 + \alpha_s(t)$ with $\alpha_0 \sim N(\dot{a}, \tau_0^2)$, and $\alpha_s \sim GP(\dot{b}, K_{\boldsymbol{\tau}})$, so that

$$\mu(t) = \dot{a} + \dot{b}(t) + X_0 + X_1(t) + X_2(t) + X_3(t), \tag{28}$$

with $X_0 \sim N(0, \tau_0^2)$ and $X_l(t) \overset{ind}{\sim} GP(0, \tau_l^2 K_l)$, $l = 1, 2, 3$.

The first eigenvalue, eigenfunction pair for $\tau_1^2 K_1$ are trivially $(\tau_1^2, \mathcal{P}_1)$, with the rest zeros. The same is true for $\tau_2^2 K_2$ with $(\tau_2^2, \mathcal{P}_2)$ and the rest zeros. The number of non-degenerate eigenpairs for $\tau_3^2 K_3$ is not finite, and they do not have a convenient closed form, but they

can be approximated with a single eigen-decomposition of the matrix $\boldsymbol{K}_3$ resulting from evaluating K_3 on a dense tensor product grid (e.g., of $M = 300$ equally spaced points) $\boldsymbol{t} = [t_1, \ldots, t_M]'$ on (-80, 80) in our case. The resulting eigen-decomposition $(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\Psi}})$ of the matrix $\boldsymbol{K}_3$, provides an approximation to the eigenvalue, eigenfunction pairs of the K_3 covariance function (e.g., see Ramsay & Silverman (2005), pp. 161, 165). That is, the k'th column of $\tilde{\boldsymbol{\Psi}}$ is an approximation to the k'th eigenfunction ψ_k of K_3 evaluated at each point in $\boldsymbol{t}$, while the corresponding k'th element of $\tilde{\boldsymbol{\lambda}}$ is an approximation to the k'th eigenvalue λ_k of K_3. Denote these approximations as $(\tilde{\lambda}_k, \tilde{\psi}_k)$, $k = 1, \ldots, M$. The error of the approximation for a particular (λ_k, ψ_k) decreases as M increases. To evaluate $\psi_k(t)$ for a point t not in $\boldsymbol{t}$, linear interpolation can be used.

Now applying the Karhunen-Loéve approximation in (25) to each X_l in (28) we obtain

$$\mu(t) \approx \dot{a} + \dot{b}(t) + b_0 + b_1\mathcal{P}_1(t) + b_2\mathcal{P}_2(t) + \sum_{k=1}^{N} b_{k+2}\tilde{\psi}_k(t), \tag{29}$$

with

$$b_k \overset{ind}{\sim} \begin{cases} N(0, \tau_k^2), & k = 0, 1, 2 \\ N(0, \tau_3^2\tilde{\lambda}_{k-2}), & k = 3, \ldots, N+2. \end{cases} \tag{30}$$

For practical purposes it is important to have $M \gg N$, and in the preceding examples $M = 300$ and $N = 50$. The GP model actually used then is technically that in (29) which is simply a Bayesian linear model.

Finally, the random effect GP curves $\nu_i(t) = \beta_{0,i} + \beta_{s,i}(t)$ from the hierarchical model in (5) follow the same GP model structure as $\mu(t) = \alpha_0 + \alpha_s(t)$ with variance parameters τ_l^2 replaced with ς_l^2. Pulling the results together, the model for each individual curve from (5) can be written as

$$y_{i,j} \approx \dot{a} + \dot{b}(\phi_{i,j}) + c_{i,0} + c_{i,1}\mathcal{P}_1(\phi_{i,j}) + c_{i,2}\mathcal{P}_2(\phi_{i,j}) + \sum_{k=1}^{N} c_{i,k+2}\tilde{\psi}_k(\phi_{i,j}) + e_{i,j}, \tag{31}$$

for $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, where

$$c_{i,k} \overset{ind}{\sim} \begin{cases} N(b_k, \varsigma_k^2), & k = 0, 1, 2 \\ N(b_k, \varsigma_3^2 \tilde{\lambda}_{k-2}), & k = 3, \ldots, N+2. \end{cases}, \tag{32}$$

and the b_k distributed as in (30). That is, the model for $y_{i,j}$ has now been reduced to a Bayesian random effects linear model for computational purposes, for which an efficient Gibbs sampling scheme can be easily implemented (see Zeger & Karim (1991), for example). Full conditionals for all parameters are closed form conjugate distributions. A major advantage to this computational scheme over MCMC algorithms used for GP models with traditional spatial covariance functions, is that a large matrix solve is never required for multivariate normal sampling or likelihood calculation inside of the MCMC. The only large matrix solve required for this algorithm is that for the $M \times M$ matrix to obtain the eigen-decomposition of K_3, which is performed just one time prior to the MCMC iterations.